THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.25, NO.4E 2006, 12 pp. 154~158

Efficient Noise Estimation for Speech Enhancement in Wavelet Packet Transform

Sung-il Jung*, Sung-il Yang*
*School of Electrical and Computer Engineering, Hanyang University
(Received October 26; Revised November 20 2006; accepted November 29 2006)

Abstract

In this paper, we suggest a noise estimation method for speech enhancement in nonstationary noisy environments. The proposed method consists of the following two main processes. First, in order to receive fewer affect of variable signals, a best fitting regression line is used, which is obtained by applying a least squares method to coefficient magnitudes in a node with a uniform wavelet packet transform. Next, in order to update the noise estimation efficiently, a differential forgetting factor and a correlation coefficient per subband are used, where subband is employed for applying the weighted value according to the change of signals. In particular, this method has the ability to update the noise estimation by using the estimated noise at the previous frame only, without utilizing the statistical information of long past frames and explicit nonspeech frames by voice activity detector. In objective assessments, it was observed that the performance of the proposed method was better than that of the compared (minima controlled recursive averaging, weighted average) methods. Furthermore, the method showed a reliable result even at low SNR.

Keywords: *Noise estimation, Speech enhancement, Best fitting regression line, Uniform wavelet packet transform, Differential forgetting factor, Correlation coefficient.*

I. Introduction

Noise Estimation is a crucial step for many speech enhancement algorithms on the single channel where speech coexists with noise. If the noise estimation is lower than pure noise, perceptually annoying musical tone may be remained. However, if it is higher than pure noise, speech distortion may be perceived. In fact, in nonstationary noisy environments, it is very difficult work to estimate the noise accurately without introducing the speech distortion and the musical tone. General methods for the noise estimation use the statistical information of nonspeech frames that are detected by voice activity detector (VAD) [1]-[3]. However, if background noise is nonstationary and speech has a weak component or is at low SNR, it may be difficult to expect the VAD to be reliable. Consequently, under various noise-level conditions, the noise estimation using the VAD may not be satisfactory [4]-[10].

In order to update the noise estimation continuously, without using the nonspeech frames by the VAD, various approaches have been proposed. For examples, minimum statistics (MS) [4] is an algorithm for updating the noise spectrum by tracking the minimum value extracted from a smoothed power spectrum of the noisy speech over a search window. The shortcoming of the MS algorithmis the requirement of statistical information of long past frames [6]-[8]. Thus, to overcome this limitation, some modified versions based on the MS algorithm were proposed in [5]-[9]. Like the MS algorithm, these modified versions update the noise estimation based on

Corresponding author: Sung-il Yang (syang@hanyang.ac.kr)
Hanyang University, Department of Computer and Electrical Eng., Sa-Dong, Sangrok-Gu, Ansan, Kyunggi-Do, Korea.

past frames required in each algorithm. A weighted average (WA) [10] is a technique to estimate the noise by applying a fixed forgetting factor between the spectrum magnitude of noisy speech at the present frame and that of the estimated noise at the previous one. However, the WA technique is difficult to estimate the noise reliably by using a fixed forgetting factor which does not consider the variation of noise in nonstationary noisy speech.

In this paper, we present an efficient noise estimation method by using the estimated noise at the previous frame only, without employing the statistical information of long past frames and nonspeech frames obtained by the VAD. First, in order to receive fewer affect of variable signals, the best fitting regression line (BFRL) [11] is used, which is calculated by applying the least squares method to coefficient magnitudes in the node with the uniform wavelet packet transform (UWPT) [13]. Next, in order to update the noise estimation efficiently, the differential forgetting factor and the correlation coefficient per subband are used, where subband is employed for applying the weighted value according to the change of signals.

II. Proposed Noise Estimation Method

Let $s(n)$ and $w(n)$ denote the clean speech signal and the noise, respectively. The noisy speech $x(n)$ is given by $x(n)=s(n)+w(n)$. The coefficient of uniform wavelet packet transform (CUWPT) $X_{i,j}^k(m)$ [13] for $x(n)$ can be expressed as

$$X_{i,j}^k(m) = S_{i,j}^k(m) + W_{i,j}^k(m) \tag{1}$$

where i, j $(0 \le j \le 2^{K-k}-1)$, k $(0 \le k \le K)$, and m are the frame index, node index, tree depth in total tree depth K, and the coefficient bin index in each node, respectively. Also, $S_{i,j}^k(m)$ is the CUWPT of the clean speech and $W_{i,j}^k(m)$ is that of the noise.

Usually, it is very difficult to execute the noise estimation accurately from the CUWPT $X_{i,j}^k(m)$ corrupted by nonstationary noise on the single channel. Thus, in order to receive fewer affect of variable signals, the BFRL $\overline{\mathbf{X}}_{i,j}^k = \left[\overline{X_{i,j}^k}(0),\ldots,\overline{X_{i,j}^k}(N-1)\right]^{\mathrm{T}}$ is used, which is calculated by the least squares method in (2), where $\overline{X_{i,j}^k}(m)$ and N are the BFRL coefficient and the node size, respectively.

$$\overline{\mathbf{X}}_{\mathrm{i,j}}^{\mathrm{k}} = \mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}|\mathbf{X}_{\mathrm{i,j}}^{\mathrm{k}}| \tag{2}$$

where $|\mathbf{X}_{i,j}^k| = \left[|X_{i,j}^k(0)|,\ldots,|X_{i,j}^k(N-1)|\right]^{\mathrm{T}}$ and $\mathbf{A}$ [11] are the magnitude coefficients in node with the UWPT and the transformation matrix of $N\times 2$ to obtain the BFRL, respectively.

The proposed noise estimation is given by (3) through (5). This method uses the differential forgetting factor $\gamma_i(\tau)$ and the correlation coefficient $\eta_i(\tau)$ obtained per subband, where the subband consists of several nodes with the UWPT. $\gamma_i(\tau)$ is the weighted value which is decided according to the amount of the noise within the subband. $\eta_i(\tau)$ is the indicator to distinguish whether it is a speech-like or a noise-like subband. $\gamma_i(\tau)$, $\eta_i(\tau)$, and $\widehat{W_{i,j}^k}(m)$ are defined as follows:

$$\gamma_i(\tau) = \min\left(1, \frac{\sum_{m=\mathrm{SB}\tau}^{\mathrm{SB}(\tau+1)} \widehat{W_{i-1,j}^k}(m)}{\sum_{m=\mathrm{SB}\tau}^{\mathrm{SB}(\tau+1)} \overline{X_{i,j}^k}(m)}\right) \tag{3}$$

$$\eta_i(\tau) = \frac{\sum_{m=\mathrm{SB}\tau}^{\mathrm{SB}(\tau+1)} (\overline{X_{i,j}^k}(m) - \mu_{\overline{\mathbf{X}}_i}(\tau))(\widehat{W_{i-1,j}^k}(m) - \mu_{\widehat{\mathbf{W}}_{i-1}}(\tau))}{\sigma_{\overline{\mathbf{X}}_i}(\tau)\sigma_{\widehat{\mathbf{W}}_{i-1}}(\tau)} \tag{4}$$

$$\widehat{W_{i,j}^k}(m) = \begin{cases} \kappa\gamma_i(\tau)\overline{X_{i,j}^k}(m) + (1-\kappa\gamma_i(\tau))\widehat{W_{i-1,j}^k}(m), \\ \quad \text{if } \gamma_i(\tau) > \mathrm{Th}_\gamma \text{ AND } \eta_i(\tau) > \mathrm{Th}_\eta \\ \widehat{W_{i-1,j}^k}(m), \text{otherwise} \end{cases} \tag{5}$$

where SB means the subband size and is equal to $2^p N$ and 2^p $(k \le p)$ is the nodes batch divided into nodes 2^{K-k} on tree depth k. Also, τ $(0 \le \tau \le 2^{K-p}-1)$, κ, Th_γ, and Th_η are the subband index, additive weight of $\gamma_i(\tau)$, threshold of $\gamma_i(\tau)$, and that of $\eta_i(\tau)$, respectively. $\mu_{\overline{\mathbf{X}}_i}(\tau)$ $(\mu_{\widehat{\mathbf{W}}_{i-1}}(\tau))$ and $\sigma_{\overline{\mathbf{X}}_i}(\tau)$ $(\sigma_{\widehat{\mathbf{W}}_{i-1}}(\tau))$ are the mean of $\overline{X_{i,j}^k}(m)$ $(\widehat{W_{i-1,j}^k}(m))$ and the standard variance of $\overline{X_{i,j}^k}(m)$ $(\widehat{W_{i-1,j}^k}(m))$ within the subband, respectively. It should be noted that there are three major differences between the proposed method and that in [10]: 1) it uses the BFRL to estimate efficiently the noise receiving fewer affect of the variable signals; 2) it applies the differential forgetting factor in noise-like

subband rather than the fixed forgetting factor in the frame; 3) it uses the correlation coefficient to discriminate the signal ingredient within the subband.

In order to get the CUWPT $\widehat{S}_{i,j}^{k}(m)$ of the enhanced speech, the modified spectral subtraction method based on the BFRL is used as follows:

$$\widehat{S}_{i,j}^{k}(m)=\begin{cases}\left(1-\dfrac{\widehat{W}_{i,j}^{k}(m)}{\overline{X_{i,j}^{k}(m)}}\right)X_{i,j}^{k}(m), \text{if } \overline{X_{i,j}^{k}(m)} > \widehat{W}_{i,j}^{k}(m) \\ \left(\lambda\dfrac{\widehat{W}_{i,j}^{k}(m)}{\overline{X_{i,j}^{k}(m)}}\right)X_{i,j}^{k}(m), \text{otherwise}\end{cases} \quad (6)$$

where λ is the spectral flooring factor [14].

III. Experimental Results

The performance of the proposed noise estimation method is evaluated and compared with that of other methods presented in [6] (MCRA: minima controlled recursive averaging) and [10] (WA: weighted average, forgetting factor $\alpha=0.95$, threshold $\beta=2$). For the implementation of the proposed method, the following parameters are chosen: frame size is 512 samples (3.2 ms) with 50% overlapping windows; the UWPT is achieved by the total tree depth $K=6$ using a Daubechies basis; $p=3$; $\kappa=0.25$; $\mathrm{Th}_{\gamma}=0.75$; $\mathrm{Th}_{\eta}=0.75$; $\lambda=0.05$. For the experiment, 20 speech signals consisting of 10 female and 10 male speakers are chosen from TIMIT database, and 2 types of noise are selected from NoiseX-92 database: speech-like noise and aircraft cockpit noise. Each speech is corrupted by each type noises with SNR 0 dB, 5 dB, and 10 dB. The sampling frequency is 16 kHz.

In order to check the SNR improvement of speech enhanced by the proposed and compared noise estimation methods, we measure the improved segmental SNR ($\mathrm{Seg.SNR}_{\mathrm{imp.}}$) as follows:

$$\mathrm{Seg.SNR}_{\mathrm{imp.}}=\mathrm{Seg.SNR}_{\mathrm{Output}}-\mathrm{Seg.SNR}_{\mathrm{Input}} \quad (7)$$

where $\mathrm{Seg.SNR}_{\mathrm{Output}}$ and $\mathrm{Seg.SNR}_{\mathrm{Input}}$ are the segmental SNR [12] of the enhanced speech and that of the noisy speech, respectively. Figure 1 is to represent the performance variations of $\mathrm{Seg.SNR}_{\mathrm{imp.}}$ according to the column dimension variation ($N\times 2$ (BFRL), $N\times 3$ (least-squares parabola),$\cdots$, $N\times N$ (least-squares $N-1^{\mathrm{th}}$ dimension polynomials)) [11] of the transformation matrix A to extract the best fitting curve. As the column dimension of the transform matrix A increases sequentially, those performances are reduced sequentially. It is the main reason why the best fitting curve for the increased column dimension has more variable than that for the former dimension. Thus, the performance for the proposed recursive noise estimator is reduced gradually by getting more influence from variable signals. The average performance difference between the BFRL and the least-squares $N-1^{\mathrm{th}}$ dimension polynomials indicates 3.03 dB in the $\mathrm{Seg.SNR}_{\mathrm{imp.}}$. Next, among objective speech quality measures, the weighted spectral slope measure (WSSM) [12] is employed, which may possess the highest correlation with subjective speech quality. The WSSM represents the smoothness degree of short-time speech spectrum, based on critical band of auditory model [12], [14], and provides the spectral distance which indicates perceptually meaningful frequency weighting [12]. We measure the segmental WSSM (Seg.WSSM) for all the frames as follows:

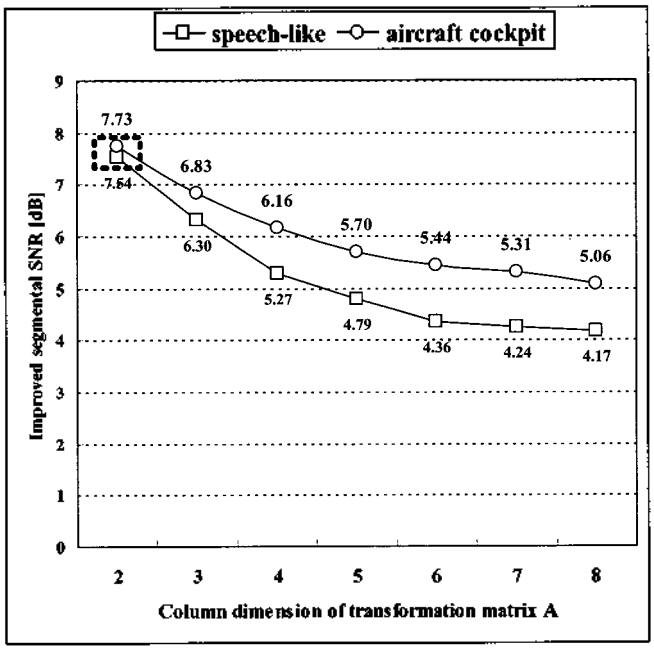

Fig. 1. Performance variation of improved segmental SNR according to the column dimension variation of transformation matrix A to extract the best fitting curve in the proposed method.

Table 1. Comparison of improved segmental SNR: WA (weighted average method [10]), MCRA (minima controlled recursive averaging method [6]), PM (proposed method).

Noise Type	Input SNR	WA	MCRA	PM
Speech-like	0 dB	6.08	7.35	**9.54**
	5 dB	5.23	6.15	**7.71**
	10 dB	4.26	4.81	**5.36**
Aircraft cockpit	0 dB	7.23	8.77	**9.54**
	5 dB	6.43	7.56	**7.74**
	10 dB	5.49	6.15	**5.92**

Table 2. Comparison of segmental weighted spectral measure: WA (weighted average method [10]), MCRA (minima controlled recursive averaging method [6]), PM (proposed method).

Noise Type	Input SNR	WA	MCRA	PM
Speech-like	0 dB	98.33	101.92	**88.97**
	5 dB	79.98	84.31	**73.48**
	10 dB	62.68	68.00	**59.35**
Aircraft cockpit	0 dB	92.05	93.37	**89.01**
	5 dB	75.30	76.49	**70.98**
	10 dB	58.95	61.33	**57.38**

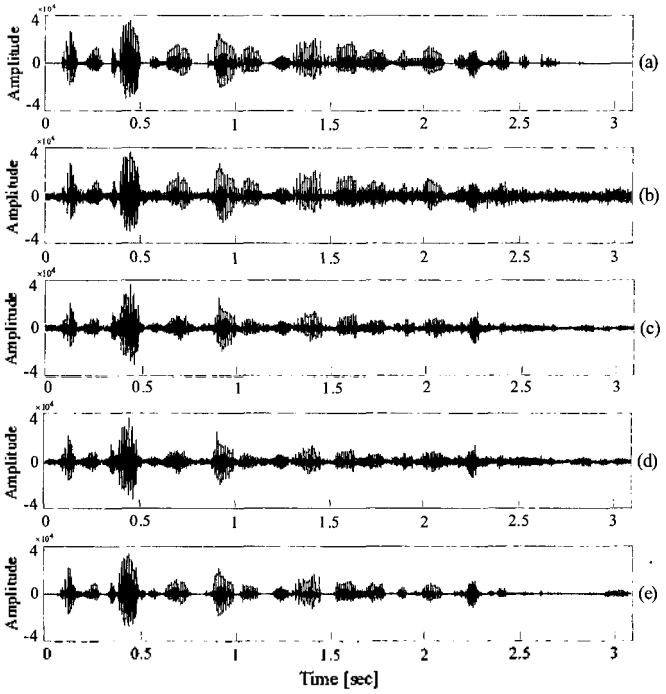

Fig. 2. Waveform of (a) clean speech, (b) noisy speech corrupted by speech-like noise at SNR 5 dB, (c) enhanced speech by the weighted average method [10], (d) enhanced speech by the *minimum* controlled recursive averaging method [6], and (e) enhanced speech by the proposed method.

$$\text{Seg.WSSM}=\frac{1}{F}\sum_{i=0}^{F-1}\left[M_{\text{SPL}}(M-\hat{M})+\sum_{r=0}^{CB-1}\Lambda_i(r)(|S_i(r)|-|\hat{S}_i(r)|)^2\right] \quad (8)$$

where F, CB, M, and $\hat{M}$ are the frame number, total number of critical band, sound pressure level (SPL) of clean speech, and the SPL of enhanced speech, respectively. Also, M_{SPL} is the variable coefficient to adjust total performance, and $\Lambda_i(r)$ is the weighting of each band.

Table 1 (Table 2) shows the $\text{Seg.SNR}_{\text{imp.}}$ (Seg.WSSM) measured from the enhanced speech by the proposed method and the compared ones. In the total average $\text{Seg.SNR}_{\text{imp.}}$ (Seg.WSSM), the proposed method has better performance 1.85 dB (4.68) than the WA method and 0.84 dB (7.71) than the MCRA one. Especially, our method shows relatively good results in more varying speech-like noise compared to other one. Consequently, the proposed method not only overcomes some drawbacks of the WA method and the MCRA method but also shows better performance than the compared methods. In Figure 2, it is observed that the enhanced speech by the proposed noise estimation method is better natural waveform than that of the compared ones

IV. Conclusions

In this paper, we proposed the noise estimation method using the BFRL, correlation coefficient, and the differential forgetting factor. Our method has the following advantages: 1) it estimates the noise efficiently using the BFRL in order to receive fewer affect of variable signals, 2) it estimate the noise adaptively using the estimated noise at previous frame only, without introducing the statistical information of long past frames and nonspeech frames obtained by the VAD. Our method showed better performance than the compared ones in objective assessments.

References

1. B. Carnero and A. Drygajlo, "Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithm," IEEE Trans. Signal Processing, **47**(6) 1622-1635, Jun. 1999.

2. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoustic Speech Signal Processing, vol. ASSP-27, **2** 113-120, Apr. 1979.
3. N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," IEEE Trans. Speech and Audio Processing, **7**(2) 126-137, Mar. 1999.
4. R. Martin, "Spectral subtraction based on minimum statistics, "EUROSPEECH, 1182-1185, Sept. 1994.
5. G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," EUROSPEECH, 1513-1516, Sept. 1995.
6. I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," IEEE Signal Process. Lett., **9**(1) 12-15, Jan. 2002.
7. I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," IEEE Trans. Speech and Audio Processing, **11**(5) 466-475, Sept. 2003.
8. S. Rangachari, P. C. Loizou, and Y. Hu, "A noise estimation algorithm with rapid adaptation for highly non-stationary environments," IEEE ICASSP, 305-308, May 2004.
9. Z. Lin and R. Goubran, "Instant noise estimation using Fourier transform of AMDF and variable start minima search," IEEE ICASSP, 161-164, Mar. 2005.
10. H. G. Hirsh and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," IEEE ICASSP, 153-156, May 1995.
11. T. K. Moon and W. C. Stirling, *Mathematical methods and algorithms for signal processing, Upper Saddle River,* (NJ: Prentice-Hall, 2000)
12. J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals, Englewood Cliffs,* (NJ: Prentice-Hall, 1993)
13. S. Mallat, *A wavelet tour of signal processing,* (2nd Ed., Academic Press, 1999)
14. N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," IEEE Trans. Speech Audio Processing, **7**(2) 126-137, Mar. 1999.

[Profile]

• **Sung-il Jung**

Sung-il Jung was born in Busan, Korea in 1972. He received the B.S. and M.S degrees in Computer Engineering from Korea Maritime University, Korea in 2000 and 2002, respectively. Currently, he is graduate student for Ph. D degree in Electrical and Computer Engineering at Hanyang University. His current research interests include speech enhancement, speech recognition, wavelet, and human robot manufacture. He is also a member IEEE, IEICE and the Acoustical Society of Korea.

• **Sung-il Yang**

He received his B.S. degree in Electronics Engineering with the greatest honors from Hanyang University, Seoul, Korea, 1984, and his M.S. and Ph.D. degrees in Electrical & Computer Engineering from the University of Texas, Austin, Texas, 1986 and 1989, respectively. Since 1990, he has been with Hanyang University and he is now a Professor at the School of Electrical & Computer Engineering.